경영학 석사 학위논문

Modelling search and choice decisions under limited product information with unstructured data

비정형데이터가 있는 제한적인 상품정보 제공환경에서의 검색과 구매 행동에 관한 연구

2020 년 2월

서울대학교 대학원

경영학과 마케팅전공

송철호

Modelling search and choice decisions under limited product information with unstructured data

지도 교수 송인성

이 논문을 경영학 석사 학위논문으로 제출함
2019년 12월

서울대학교 대학원
경영학과 마케팅전공
송철호

송철호의 경영학 석사 학위논문을 인준함
2019년 12월

위 원 장	김 상 훈	(인)
부위원장	김 병 도	(인)
위 원	송 인 성	(인)

Modelling search and choice decisions under limited product information with unstructured data

Cheolho Song
Business Administration (Marketing Major)
The Graduate School
Seoul National University

Abstract

I develop an empirical model of search and choice in which consumers are presented with limited product information prior to the search. In the model, consumers search and click on the items listed on product listing pages. They expect to view vertical as well as horizontal attribute values that cannot be observed on product listing pages (i.e. costly attribute values) after clicking-through. Vertical costly attributes include quantified review scores of several product attributes. They reflect actual users' satisfaction with the product attributes. This paper has the following contributions to the literature. First, the model reflects consumers' higher uncertainty of their utility prior to search which can be reduced by obtaining information about costly attribute values. It is in line with consumer learning literature. Second, the model also reflects consumers' heteroskedastic uncertainty of their utility during searching for the products without violating the parsimony of the model. Third, this paper uses a deep learning method in order to extract the structured features from reviews.

The model is applied to the aggregate search and choice data from Chrome-OS laptops at Bestbuy.com. The model shows the realistic values of parameter estimates and better in-sample fit in comparison with Kim et al. (2016). With the estimated model parameters, I conduct the counterfactual experiment that shows how consumer search set size and manufacturer market share and revenue change in a full information environment. In the full information environment, consumers reduce their search set size by -3.9% and choose almost the same products as they do in the limited information environment. It leads to an increase in consumer surplus by 3.19%. For producers, most of their market share and revenue increase. Furthermore, the brands with relatively low rank in total rating and high rank in average review score shows the relative higher increase. Therefore, I want to suggest to manufacturers that they should post quantified review scores with respect to each attribute on product listing pages in order to boost their sales and revenues especially when their total rating is relatively low.

Keywords: consumer search, consumer learning, limited information provision, Convolutional Neural Network, review data, durable goods

Student Number: 2018-29680

Table of Contents

Table Index

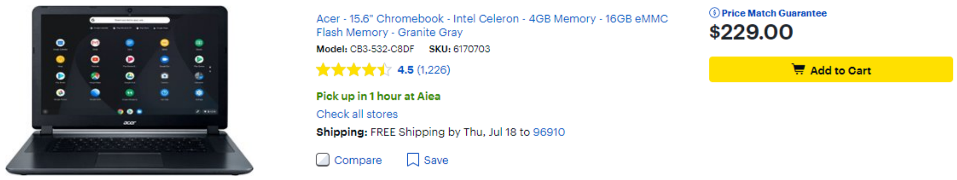
Figure Index

1 Introduction

The consumer search behavior has recently been recognized as an important topic in marketing and economic research for the following reasons. First, consumers do not consider the universal consideration set for the reasons such as nonzero search cost and consumer’ s cognitive limitation that blocks consumers from remembering all products’ information. Therefore, if the search behavior is not reflected in the model, the bias of estimates necessarily occurs because of endogeneity between the choice decision and limited consideration set. Second, the consumer search data set reveals consumer preferences (Kim et al. 2010) as the choice data sets have done in traditional marketing literature.

The online consumers’ search and choice data have been available in the form of clickstream data, a number of which show variations in individual level(Montgomery et al., 2004; Chen & Yao, 2016) and in the form of aggregate product-level search and choice data (Kim et al., 2010; Kim et al., 2016). The common assumption among these models is that consumers already know all vertical product attribute values prior to the search, and their purpose of the search is to find the horizontal product match values. This assumption is reasonable for some empirical contexts, but it must not be suitable in other online retailing environments where consumers have to move from product listing page to product detail pages in order to be fully aware of the vertical product attribute values as well as horizontal match values. For example, the renown online retailers, Amazon and Bestbuy.com, do not reveal all vertical product attribute values on the product listing page for some categories of products and therefore, it should be unreasonable to assume that consumers are fully informed with them before clicking on the product detail page.

Figure 1. Bestbuy product listing page



Acer - 15.6" Chromebook - Intel Celeron - 4GB Memory - 16GB eMMC Flash Memory - Granite Gray

Model: CB3-532-C8DF **SKU:** 6170703

★★★★☆ **4.5** (1,226)

Pick up in 1 hour at Aiea

Check all stores

Shipping: FREE Shipping by Thu, Jul 18 to 96910

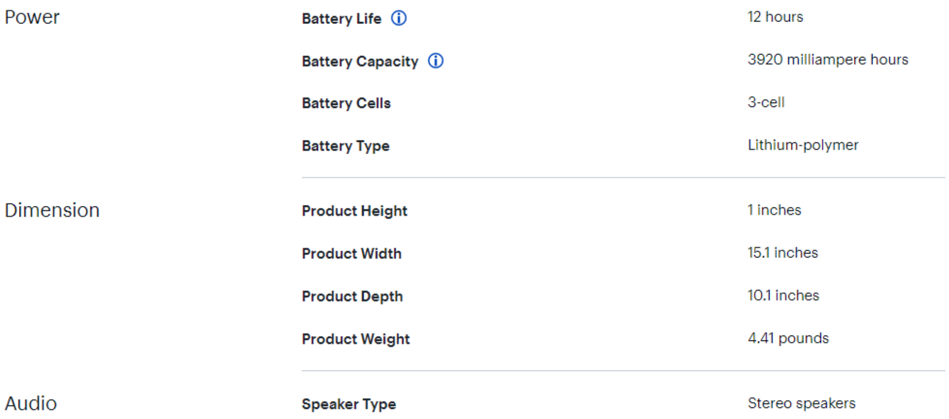
□ Compare Save

Price Match Guarantee

$229.00

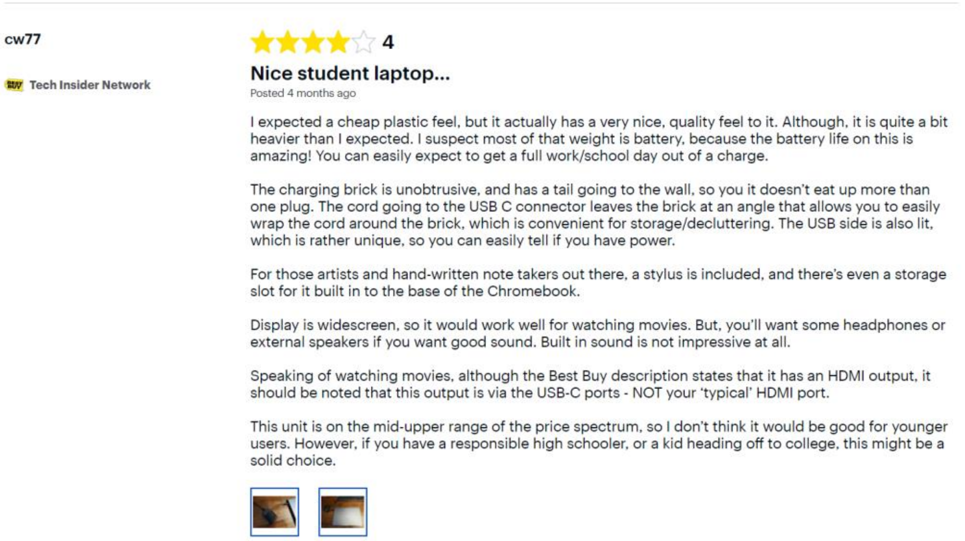
Add to Cart

Figure 2. Product detail information



Power	**Battery Life**	12 hours
	Battery Capacity	3920 milliampere hours
	Battery Cells	3-cell
	Battery Type	Lithium-polymer
Dimension	**Product Height**	1 inches
	Product Width	15.1 inches
	Product Depth	10.1 inches
	Product Weight	4.41 pounds
Audio	**Speaker Type**	Stereo speakers

Figure 3. Product reviews



cw77

Tech Insider Network

★★★★☆ 4

Nice student laptop...

Posted 4 months ago

I expected a cheap plastic feel, but it actually has a very nice, quality feel to it. Although, it is quite a bit heavier than I expected. I suspect most of that weight is battery, because the battery life on this is amazing! You can easily expect to get a full work/school day out of a charge.

The charging brick is unobtrusive, and has a tail going to the wall, so you it doesn't eat up more than one plug. The cord going to the USB C connector leaves the brick at an angle that allows you to easily wrap the cord around the brick, which is convenient for storage/decluttering. The USB side is also lit, which is rather unique, so you can easily tell if you have power.

For those artists and hand-written note takers out there, a stylus is included, and there's even a storage slot for it built in to the base of the Chromebook.

Display is widescreen, so it would work well for watching movies. But, you'll want some headphones or external speakers if you want good sound. Built in sound is not impressive at all.

Speaking of watching movies, although the Best Buy description states that it has an HDMI output, it should be noted that this output is via the USB-C ports - NOT your 'typical' HDMI port.

This unit is on the mid-upper range of the price spectrum, so I don't think it would be good for younger users. However, if you have a responsible high schooler, or a kid heading off to college, this might be a solid choice.

For example, many online retailing platforms reveal product information in several stage structure. As you can see from Figure1, the laptop category in Bestbuy provides the values of some vertical attributes on the product listing page. The detail of the product and its actual users' reviews can be accessed by clicking through the product detail page. This page includes the values of vertical attributes (Figure2) and also users' reviews (Figure3), both of which are not posted on the product link. Therefore, I cast doubts on the validity of the assumption that consumers are fully aware of vertical attributes before searching. Ghose et al. (2018)., Choi & Mela (2016) and Gardete & Megan Antill (2019) reflect the limited information environment that provides product information in multiple stages. Especially, Ghose et al. (2018) quantify users' reviews and use them as vertical attributes while several pieces of research (Kim et al., 2010; Kim et al., 2016; Chen & Yao, 2016;

Ursu, 2018) reflect them to idiosyncratic match values in the model.

In this paper, I adopt the sequential search model of Weitzman framework(1979), the seminal theoretical approach which has been recently adopted in empirical context(Kim et al, 2010; Kim et al, 2016; Chen & Yao, 2016; Ursu, 2018; Ghose, 2018). The Weitzman-based empirical sequential search model is superb one in the sense that it dramatically reduces the computational burden of solving the optimal stopping problem of search sequences by using the concept, "search cost" and "reservation utility". Moreover, previous researches handling highly differentiated durable goods in search models adopt sequential search strategies since there exists such a huge number of alternatives that it is unreasonable to assume that consumers decide what and how many products to include in their consideration sets prior to the search. The fixed-sample strategy proposed by Stigler(1961) is adopted in the research studying the market where the number of alternatives is limited (eg. car insurance market) and consumers are uncertain about only a few attributes such as price (Honka, 2017).

However, unlike previous researches of the sequential search model, I relax an assumption and let the model take it into consideration that consumers observe only a subset of vertical attributes on the product listing page (hereafter, "costless attributes"). By clicking through the detail page and paying search cost, they can find the value of the rest of the vertical attributes (hereafter, "costly attributes") including quantified review scores as well as horizontal match values. Under the limited product information environment, I assume that consumers form an expectation of the unknown values of the costly attributes conditional on the costless attribute values prior to the search. Then they construct the search sequence based on the values of costless attributes and conditional expectations of costly attributes. While searching, they 'learn' the true values of costly attributes and after finishing searching, they choose the product that they want to buy.

This paper seeks to contribute to the empirical search literature in several ways. First, the proposed model maintains the parsimoniousness of the Weitzman–based search model although the model captures flexible consumer behaviors during the searching and purchasing phase. To be specific, the model reflects consumers' different sensitivities to some subset of attributes in search and choice stages because of the uncertainty about costly attributes but their search and choice decisions are based on the identical utility. That is, their preferences are uniform during searching and purchasing stage. This setting allows both the consistency and flexibility of consumer behaviors to be held in the model. Thus, permitting flexibility does not violate the assumption of consumer' s rationality. It contrasts with past research on consideration set that explains the different sensitivities by adopting distinct utility components between search and choice stages(Moe, 2006). In our setting, the different sensitivity is explained by different information sets available during search and choice. Therefore, consumers can learn more about the products by searching them and then have more information sets, which makes consumers have less uncertainty about the products. This mechanism is in line with that of consumer experiential learning literature (Ching et al., 2013; Erdem et al., 1996).

Second, our model introduces heteroskedastic utility variance into the search stage. Compared with Kim et al. (2016), Ursu(2018) and Ghose(2018), which assume identical search utility variance among consumers, utility variances in our approach differ across consumers. Roughly speaking, such heteroscedasticity is driven by the uncertainty of the costly attributes and heterogeneous consumer preferences. Consumers who have stronger preferences for unknown attributes are more likely to search items of which costly attributes are expected to have larger variations. It corresponds to the argument of literature both from consumer learning and search topics. Consumers have a higher incentive to learn or search the items of which they are more uncertain about the quality (Erdem et al, 1996). Therefore, it makes sense that those who have a higher preference for costly attributes are more likely to search and learn about them.

Lastly, this paper extracts quantified features from unstructured text data by using a deep learning model. Despite the growing popularity of deep learning, a few pieces of research utilize them to apply to a marketing context. Some researches applied deep learning models mainly for the purpose of extracting features. However, their applications are limited to the reduced-form approaches in Marketing (Liu, 2017; Liu, 2018). In the consumer search literature, Ghose et al. (2018) utilize Latent Dirichlet Allocation (LDA) (Blei et al., 2003), one of popular machine learning methods in Natural Language Processing to extract the topics included in the reviews. Therefore, this paper would be one of the pioneering trials of applying deep learning models to extract features to a structural model.

For the empirical analysis, I apply the proposed model to aggregate−level consumer search and choice data of the Chromebook category at Bestbuy.com. I first describe the way that data are extracted from the webpage and features are refined from review data. Moreover, I also explain in detail how search and choice raw data are transformed into the dependent variables. Moreover, before talking about the main model, I analytically describe the intuition of consumers' different sensitivities to some attributes between searching and purchasing stages in a limited product information setting. Empirical model−free evidence is also presented to support the existence of different sensitivities. With the parameter estimates of the main model, the counterfactual analysis is also conducted. It shows that consumers have increased surplus by reducing the search set size under a full information environment. In other words, if the costly attribute information is revealed on the product listing page, they can save their efforts and time to search for the best alternative. The market share and revenue of manufacturers are also changed due to the different information provision.

The rest of the paper is organized as follows. In Section 2, I document the summary of data and the extraction procedure of some types of data. Section 3 contains the analytical intuition of the model and empirical model-free evidence for the main model. Section 4 presents the main model specification and Section 5 discusses its estimation and identification strategy. Finally, Section 6 presents and discusses the result of estimation and Section 7 shows a counterfactual analysis with the parameter estimates from Section 6.

2 Data

I utilize aggregate-level consumer search, choice, and product information data from a laptop category in Bestbuy.com. Many categories of durable goods are used in a dynamic structural modeling setting(Song and Chintagunta, 2003; Gowrisankaran and Rysman, 2012; Kim et al., 2010; Kim et al., 2016). The category in this paper is narrowed down to Chromebook, one of the types of laptops that uses Chrome as an Operation System.

I collected data for all Chromebooks from the middle of March to the middle of April 2019 on a daily basis. Data contains product specifications, users' reviews, and a list of other products that were searched or purchased by consumers who viewed the focal product and sales rank data. Then, I aggregated the time-varying data to longitudinal ones. For time-varying variables, the average price over the period and the latest reviews for products are adopted for the analysis. By aggregating the data, each product has a list of a sufficient number of other products that were browsed or purchased.

2 – 1 Details of Search and Choice Data

To clarify the search and choice data, I explain what exactly search and choice data is and how they are transformed from raw data which are essential for inference of the model in this paper. The transformed data includes the relative view rank data, conditional share, and sales rank data. In order to create relative view rank data, Kim et al. (2016) utilize the aggregate–level search data set, 'Customers who viewed this item also viewed' , from Amazon. Similarly, Bestbuy.com provides 'People also viewed' set which is an analog of a search data set from Amazon. This raw search data set from Bestbuy.com is a list of products that were viewed by past consumers, who viewed a focal product in the same browsing session. The product position in the search data in Kim et al. (2016) serves as a relative rank of products. In other words, if there are A and B products in C' s searched product list and A is located left to B, then A is more often viewed with C than B. Therefore, Kim et al. (2016) uses position–based search popularity (See Table1) to construct relative view rank data. However, there is no guarantee that the product position in Bestbuy.com search data set represents relative ranking among them but this paper uses position–based search popularity to construct view rank data. The reason for it is explained in section 3–3.

Table 1. Constructing search popularity

1) appearance-based search popularity: $SearchPopularity_{jl}$ is 1 if product j appears on product l' s view list, and 0 otherwise

2) position-based search popularity

$$SearchPopularity_{jl} = \frac{ViewListLength_l + 1 - Position_{jl}}{ViewListLength_l}$$

where $ViewListLength_l$ is the number of products that appear on l' s view rank list (in my case, six products are shown on Bestbuy' s search data list). $Position_{jl}$ is the j' s position in the l' s search data list. The value of Position is the lowest if it is located at the upper-most position and the highest if it is located at the lowest position.

I collected the search data list of all products from the Bestbuy.com Chromebook category on a daily basis. On a day, each product has six products in their search data list. I recorded and sum up (or average) the search popularity of all products in each product search data list to construct appearance- (position-) based view rank lists. Table 2 is the part of the appearance-based search data list of a product aggregated over a data collection period.[1] The row number means the focal product number and the column number is the product number in the focal product. For example, (1,2) has 7, which means the product 2 appeared seven times in the search data list of the product 1. Then I transformed

[1] I explain the construction of view rank list based on appearance-based search popularity for the convenience of explanation. The same logic is also applicable to position-based view rank list. The difference is that position-based view rank list is aggregated by averaging search popularity over time period.

this $J \times J$ matrix into $J \times J \times J$ view rank inequality matrix which is in a suitable form for the model estimation. The third dimension of view rank list represents the focal product and the others are the products in the focal product’s search data list. Each cell of this view rank inequality matrix compares the appearance frequency among products that are contained in the same focal product’s search data list. For instance, (1,2,3) in the view rank inequality matrix is 1 because the product 1 appears more often than the product 2 in the search data list of the product 3.

Table 2. Search data list (appearance base)

Products Focal products	1	2	3	4
1	0	7	4	2
2	1	0	2	6
3	1	3	0	0
4	9	11	4	0

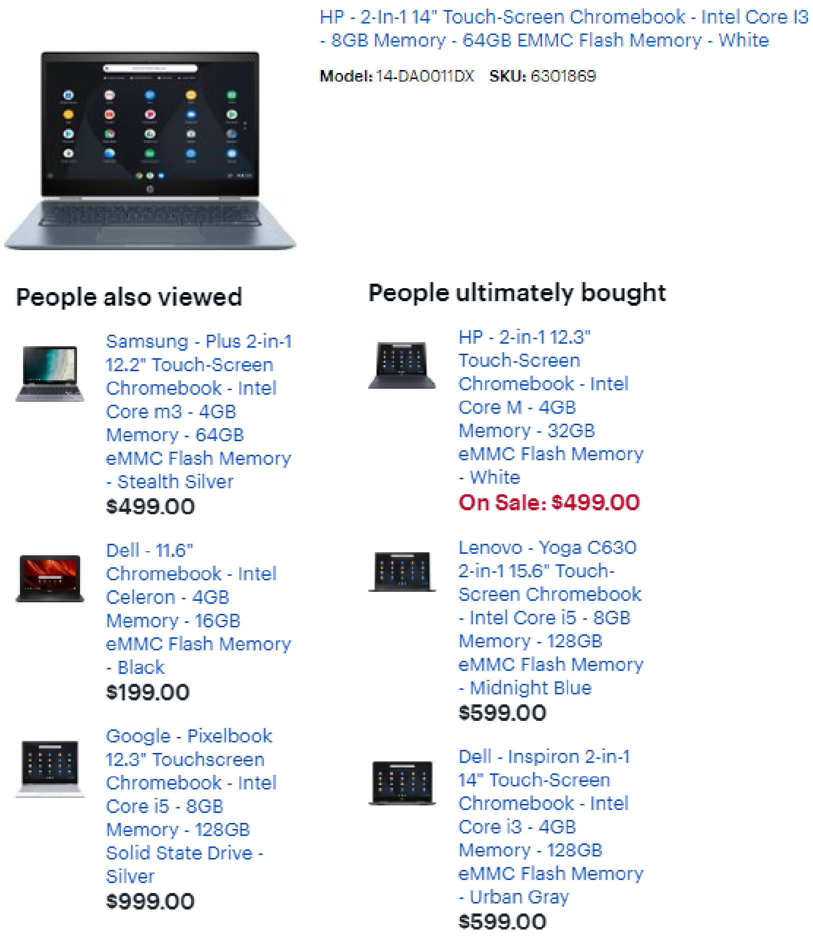
The conditional share data consists of the choice shares of products in the category, conditional on viewing a focal product. In other words, if product B is often chosen among consumers who viewed product A, B will appear often on product A’s conditional share list. The raw data of the conditional share list comes from ‘People ultimately bought’ of a focal product in Bestbuy.com. In collecting and aggregating the conditional share data, I counted the number of appearances of each product and then average them. For example, as shown in Table 2, 6.25% of those who viewed product

1 and finally decided to buy the products chose product 2.

Table 3. Conditional share data

Products Focal Product	1	2	3	4
1	0	0.0625	0.0802	0.0513
2	0.0556	0	0.0679	0.0256
3	0.0123	0.0069	0	0.0192
4	0.0617	0.0208	0.0556	0

Figure 4. Raw search data



HP - 2-In-1 14" Touch-Screen Chromebook - Intel Core I3 - 8GB Memory - 64GB EMMC Flash Memory - White

Model: 14-DA0011DX **SKU:** 6301869

People also viewed

Samsung - Plus 2-in-1 12.2" Touch-Screen Chromebook - Intel Core m3 - 4GB Memory - 64GB eMMC Flash Memory - Stealth Silver
$499.00

Dell - 11.6" Chromebook - Intel Celeron - 4GB Memory - 16GB eMMC Flash Memory - Black
$199.00

Google - Pixelbook 12.3" Touchscreen Chromebook - Intel Core i5 - 8GB Memory - 128GB Solid State Drive - Silver
$999.00

People ultimately bought

HP - 2-in-1 12.3" Touch-Screen Chromebook - Intel Core M - 4GB Memory - 32GB eMMC Flash Memory - White
On Sale: $499.00

Lenovo - Yoga C630 2-in-1 15.6" Touch-Screen Chromebook - Intel Core i5 - 8GB Memory - 128GB eMMC Flash Memory - Midnight Blue
$599.00

Dell - Inspiron 2-in-1 14" Touch-Screen Chromebook - Intel Core i3 - 4GB Memory - 128GB eMMC Flash Memory - Urban Gray
$599.00

2 – 2 Data Summary

I select the 60 Chromebooks for analysis which have both search data and conditional purchase share information and also appeared at least one time in other products' search lists and conditional share. The descriptive statistics for these products are listed in Table 3.

Table 4.Descriptive statistics of Chromebooks

Brands		Acer(17), ASUS(7), Dell(8), Google(4), HP(12), Lenovo(4), Samsung(8)
Price		$356.49(mean), $252(std. dev.)
2-in-1		Yes(27), No(33)
Screen size		13.1(mean), 1.48(std. dev.)
Storage capacity		50.1GB(mean), 72.3GB(std. dev.)
ram		4.4GB(mean), 2.10GB(std. dev.)
eMMC		Yes(53), No(7)
color Black		Yes(10), No(50)
Total rating		4.45(mean), 0.6(std. dev.)
Review #		148(mean), 299(std. dev.)
Review scores	Speed	0.12(mean), 0.14(std. dev.)
	Price	0.22(mean), 0.19(std. dev.)
	School	0.13(mean), 0.15(std. dev.)
The review scores are extracted by the pre-trained classifiers using Convolutional Neural Network		

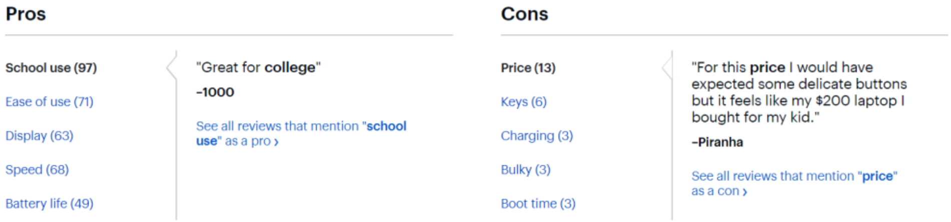
2 – 3 Review Feature Extraction

Users' reviews of products are available in Bestbuy.com if they exist. Review ratings are scored on 5 scales. The format and content of reviews look like Figure 3. Since reviews contain users' satisfaction with the products, it can be valuable information for potential buyers to decide whether to buy or not. They cannot be obtained from the product' s specifications by the retailer. For example, users' satisfaction with laptop' s speed is provided only by actual users.

However, since text data are an unstructured type of data, it is hard to quantify and obtain interpretable attributes and users' satisfaction with them from the text itself. In Machine Learning literature, there are two types of methods widely used to manipulate text into quantified data. One is Latent Dirichlet Allocation (LDA), one of the unsupervised methods that can be used without a label. The other is classification methods such as Support Vector Machine, Neural networks and so on. The latter methods are supervised methods that require labels which observations are classified into.

A number of researches were conducted, which classify text into positive or negative sentiment based on ratings. However, such a sentiment analysis based on ratings is not useful for my case. This is because it does not provide satisfaction with the product' s specific attribute if the label indicating the existence of the attributes in the review and the corresponding satisfaction does not exist. Therefore, in order to extract consumer satisfaction with the product' s attributes, I need a break-through to obtain such labels. Liu et al. (2017) employ a Neural Network model to classify reviews into Positive/Negative sentiments along with each attribute; hence, the number of classifiers is equal to the number of desired attributes. Liu et al. (2017) obtain desired labels (whether product attributes exist and if they do, whether they are considered positive or negative) by getting help from manual forces in Amazon Mturk. To be more specific, they upload reviews that are to be used as training data set onto the Mturk surveys and request Mturk users to answer whether certain product attributes are mentioned in the reviews and whether they are positive or negative. After obtaining a sufficient number of answers from Mturk, they use them to train classifiers and then the trained classifiers predict a test data’ s label, which are used in the main model. Predicted labels of each attribute are then aggregated into the product level and used as the attributes in the main model. In short, pre−trained classifiers are utilized to extract the features from the text and the product−level aggregations of the features are finally adopted as variables in the main model.

Figure 5. Training data and labels for CNN classifiers



Pros

School use (97)
Ease of use (71)
Display (63)
Speed (68)
Battery life (49)

"Great for **college**"
-1000
See all reviews that mention "**school use**" as a pro ›

Cons

Price (13)
Keys (6)
Charging (3)
Bulky (3)
Boot time (3)

"For this **price** I would have expected some delicate buttons but it feels like my $200 laptop I bought for my kid."
-Piranha
See all reviews that mention "**price**" as a con ›

I utilize a similar method to extract subjective satisfaction with each attribute. Instead of using manual forces from Mturk, the exiting attributes and sentiments are used to train classifiers. Figure5 is ‘Pros and Cons’ review lists of each attribute from Window−OS laptop which are provided from Bestbuy.com. Since products in Chrome−OS laptop category are exclusive to ones in

Window−OS laptop category, Pros and Cons review data of Window−OS laptop are appropriate to train and validate classifiers which are to be used for predicting the labels of Chrome−OS laptops (i.e. reviews from Window−OS laptop category are divided into training and validation data set for CNN estimation, and reviews from Chrome−OS laptop category are used as a test/prediction data set)[②]. Among many attributes, I select satisfaction with ‘speed’, ‘price’, ‘school use’, which are among the most frequently mentioned attributes in both Chrome−OS laptops and Window−OS laptops.

The mean(median) number of reviews per product in the training data set and test data set is 246(95) and 264.2(31), respectively. The sentiments of each attribute in the training data set are not evenly distributed. Neutral reviews are the most prevalent, positive reviews are the next, and there are a few negative reviews across all attributes. Even the ‘School use’ attribute does not have negative reviews in my data. Although I acknowledge that negative reviews could have an impact on consumers' decisions, they make it difficult to estimate the classifier because such a highly unbalanced label inhibits the model to converge and make the prediction of other labels less accurate. For this technical matter, the negative label of each attribute is omitted. Therefore, binary labels (Positive vs Neutral or negative) are adopted for every classifier.

[②] For simplicity, ‘training data set’ refers to reviews from window-OS laptop and ‘test data set’ or ‘main data set’ refers to reviews from Chrome-OS laptop.

Table 5. Distribution of review numbers per product

	1st quantile	median	mean	3rd quantile	max
Training data	15	95	246	277	3272
Test data	6	31	264.2	213	1846

Table 6. Sentiment distribution of attributes

Attributes	Negative	Neutral	Positive
Speed	120	15703	5654
Price	114	15913	5450
School use	0	19112	2365
# of observations = 21477			

Table 7. Examples of sentiment assignment on reviews

Reviews	Speed	Price	School
- Good speed - Adequate storage - Great for students and Sims 4(game) - Good for streaming	1	0	1
it was not worth it. When trying to exit out of programs or going to another webpage it takes forever. If you do light browsing and just surf the web this computer would be perfect for you. Would not recommend writing a paper or handling business-related things.	-1	0	0
1: positive, 0: neutral, -1: negative			

2-3-1 Convolutional Neural Network for Extracting Features

Convolutional Neural network (CNN) is used as the text classifying model. CNN is a popular model in computer vision and NLP researches because of its distinctive characteristics from other deep learning models. CNN can capture local clues through convolution (or local filters) and it uses a pooling method which makes the model location-insensitive.

The architecture of the models used in this research is almost the same as one of Liu et al. (2017). It has four layers and the first layer is the word embeddings of product reviews. The second layer is the convolutional layer. The third layer is the max-over-words pooling layer. In the fourth layer, all third layers are concatenated into a one-dimensional layer and the sigmoid function is applied to them so that it can be matched with a binary sentiment label. Since I want to create three features, the three separate CNN models with different attribute labels are trained.

Layer 0: review data preprocessing

Each review is regarded as one observation or a document in the NLP term. Documents are tokenized into a word and then, transformed into a sequence of integers (each integer is the index of a token in a dictionary). I padded each tokenized document with zeros next to each side of documents so that they have their length to be one of the longest reviews. So, all documents except for the longest reviews have null cells.

Layer 1: Word Embedding

Although sentences or a combination of words have semantic meaning, the preprocessed tokenized matrix does not reflect it. For example, the original review '- Good speed - Adequate storage - Great for students and Sims 4(game) - Good for streaming' is transformed into [0,0,0,…., 4, 482, 1238, 77, 63, 817, 4, 66, 462, 1, 632, 286, 275, 1957], where consecutive sequence of zeros represents a padding. One could intuitively understand by seeing this sequence that it does not reflect any semantic meaning. The review content' s information can be represented by low-dimensional pre-trained word-embeddings. We utilize the word2vec embeddings published by Google. These embeddings are trained on 100 billion words from the Google news dataset using the method of Mikolov et al. (2013). The embeddings have the words with similar context occupying close spatial positions and dissimilar words far from each other. Thus, each word is represented by a 300-dimensional vector. By using mathematical notation, i-word in the review can be represented as $\vec{x}_i \in R^k$ where k = 300 , a review can be represented as $\vec{x}_{1:N} = \vec{x}_1 \oplus \vec{x}_2 \oplus \ldots \oplus \vec{x}_N$ where N = 1700, a maximum length of reviews and $\oplus$ is the concatenate operator. Thus, one review is an Nk-dim vector.

Layer 2: Convolution Operation or Filter

In the next layer, the word embeddings from the first layer go through the convolution operation. The convolution operator is a one-dimensional vector of length h, applied to each sliding window of h words which has s strides. In this setting, h and s are set as 2 and 1, respectively. Thus, it works as a bigram filter. To be more specific, the convolution operator is a hk*1 vector where h(=2) is the window size and k(=300) is the dimensionality of the word embeddings. Let i be the current position of the convolutional operator and then, $\vec{x}_{i:i+h-1} \in R^{hk}$ be a window that the operator applied to. The output of the convolution operation is $c_i = ReLU(\overrightarrow{w} \cdot \vec{x}_{i:i+h-1} + b)$. Rectified linear units (ReLU) is the chosen for the activation function (Goodfellow et al. 2016), where the

ReLU function is defined as $\mathbf{ReLU(x)} = \mathbf{max}\ (x, 0)$. One example of alternative activation functions is the Sigmoid function, also known as a logit function. The Sigmoid function is one of the most widely used activation functions. However, its gradient can vanish at either end of the sigmoid function, which is called the "vanishing gradients" phenomenon. It makes the neural network refuse to learn further. Instead, the ReLu function does not vanish at any point as a linear function does although ReLU is nonlinear in nature. Moreover, ReLU is less computationally expensive than the Sigmoid function because of its simpler mathematical operations. Anyhow, the convolutional operator is rolled over i-th review' s embeddings where i=1,2,⋯, N. The final output is a vector $\vec{c} = [c_1, c_2, \ldots, c_{N+1}]$.

<u>Layer 3: Pooling</u>

In the third layer, the max-over-time pooling operator is applied to the feature map from the convolution layer. The feature map going through the max-over-time pooling operator brings the outcome such that $\hat{c} = \mathbf{max}\ \{\vec{c}\}$. One can understand that the outcome is the most salient information across bi-gram tokens in layer 2. In other words, $\hat{c}$ is the bi-gram representation of the whole information of a review and it captures the most indicative information in the review.

<u>Layer 4: Append and Output</u>

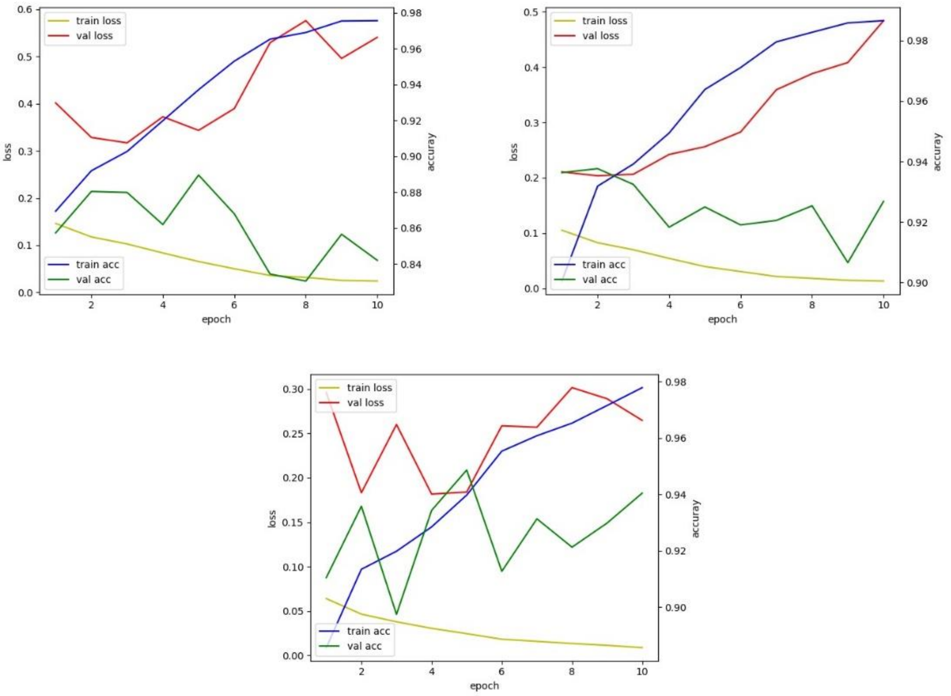
In the final layer, the outcome from the layer 3 is flattened and the sigmoid activation function is applied to it. Then, it provides the probability that the review contains the positive contents of an attribute. Using this probability, a weighted binary cross-entropy value (BCE) is calculated as follows:

$$\mathrm{BCE(y)} = -\frac{1}{J}\sum_{j=1}^{J} w y_j \log(pos_{prob_j}) + (1-w)(1-y_j)\log(1-pos_{prob_j})$$

$$where\ pos\_prob_j = sigmoid(\beta \cdot \hat{c}_j) = \frac{\exp(\beta \cdot \hat{c}_j)}{1+\exp(\beta \cdot \hat{c}_j)}\ where\ j = 1, \dots, J(= 16107)$$

where w is an adjusting weight for the imbalance of classes, and $\mathbf{y}_j$ indicates whether j–th review contains positive content. w is close to zero or one if the imbalance between classes is severe and equals to 0.5 if the number of observations from the two classes is the same. BCE is adopted as the loss function of the model.

Figure 6. CNN learning graphs



(upper left: Speed, upper right: Price, lower: School)

Upon completion of training the model, I chose the model with the largest validation accuracy, the smallest validation loss and the decent level of training accuracy in order to avoid under- and over-fitting. I use them to predict the outcomes of test data. Table 7 shows the results. There are 11454 reviews throughout all products. CNN-predicted attributes are the values indicating whether a review contains positive information related to an attribute. Then, I aggregate each attribute of reviews into product-level values by averaging them. For instance, let a 2nd product has 70 neutral or negative reviews and 30 positive reviews with respect to speed. Then, this product has an aggregated speed score of 0.3. These product-level review scores are utilized in the main model of this paper as costly attributes. I assume that consumers look through the review of products and have the average scores with regard to attributes in their mind and consumers do not face heterogeneous average scores.

One can suggest LDA as an alternative training model. Since LDA does not need to estimate the model with training data, the model was directly applied to the test data. However, from a different view, the information from training data cannot be used for LDA. I manually put labels to each review based on the probability for each topic to appear on reviews. Then, as CNN review scores are derived, I extract product-level LDA review scores. After then, for the comparison of two types of features, the sales ranks of products are regressed on LDA review scores and its result is compared with the results of regression on CNN review scores.

From Table 8 and Table 9, I conclude that CNN review scores explain better the variation of sales rank. Therefore, I abandon LDA results and adopt CNN results for choosing variables of the main model.

Table 8. CNN predictions of test data

Attributes	Neutral	Positive
Speed	9803	1651
Price	7747	3707
School use	8671	2783
# of observations = 11454		

Table 9. Regression of sales rank on LDA features

D.V: sales rank	Estimate	Std. Err
Intercept	17.10 (***)	3.61
Use	89.70*	37.00
Price	6.66	23.99
Memory	−1.45	17.31
internet	34.11	30.43
School	8.11	22.45
screen	17.16	37.50
R−squared: 0.301	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1	

Table 10. Regression of sales rank on CNN features

D.V: sales rank	Estimate	Std. Err
Intercept	18.81 (***)	3.15
Speed	52.79 (***)	13.88
Price	9.30	11.88
School	26.20 (.)	15.35
R−squared: 0.3297	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1	

3 Empirical Settings

3 – 1 Product Information Environment

The information environment of Bestbuy.com is documented in this section. Bestbuy.com provides a category page in which consumers can find a list of products with an image and summary of the product in the form of a URL links. Figure1 shows an example of a product on the product listing page. Here we can see that the link contains a subset of product information with some vertical attributes. It is reasonable to assume that consumers can learn a subset of product information by just browsing the product listing page. However, the link does not reveal information on other vertical attributes which can be observed by visiting the product detail page. Note that the links at the product listing page reveal a limited set of information whereas the product detail page provides full product information. I analytically show how consumers have their utility during searching and purchasing phases under the limited information environment. To put it briefly, consumer shows the different sensitivities to some attributes during searching and purchasing stages.

Consider utilities of search and choice such that

$$u_j^c = V_j^c + e_j^c = V_j^f + V_j^{nf} + e_j^c \text{ : purchase utility (1)}$$

$$u_j^s = \widetilde{V_j^c} + e_j^s = V_j^f + \widetilde{V_j^{nf}} + e_j^s \text{: search utility (2)}$$

$$\text{where } V_j^f = X_j^f \cdot \beta_1 \; and \; V_j^{nf} = X_j^{nf} \cdot \beta_2$$

$$\text{and } \widetilde{V_j^{nf}} = \widetilde{X_j^{nf}} \cdot \beta_2 \, .$$

In the equations (1) and (2), superscripts c and s stand for choice and search, respectively, f and nf stand for free (costless attribute available on listing pages) and not-free (costly attribute only observable on product detail pages), respectively, and $\boldsymbol{\beta}$ is consumer sensitivities to vertical attributes. Purchase utility consists of determinant parts of free (V_j^f) and not-free vertical attributes (V_j^{nf}), and horizontal idiosyncratic matching value e_j^c. The search utility is comprised of a determinant part of free attributes (V_j^f), a random part of not-free attributes ($\widetilde{V_j^{nf}}$) and horizontal idiosyncratic matching value e_j^s. Consumers know the existence of free and not-free vertical attributes but observe only free vertical attributes prior to the search. Therefore, in the searching stage, consumers form expectations of unknown vertical attributes using free attributes. I denote not-free vertical attributes as a linear specification of free vertical attributes added with a random variable υ_j in the search stage as follows:

$$\widetilde{X_j^{nf}} = X_j^f \cdot \gamma + \upsilon_j$$

where $\upsilon_{j,}$ is an error term for product j and makes not-free vertical attributes in the searching stage as random variables. The rationale for this random linear specification is that consumers are fully aware of the existence and distribution of not-free vertical attributes and expect the values of not-free attributes based on those of free attributes. Thus, consumer’s search utility can be rewritten as,

$$u_j^s = X_j^f \cdot (\beta_1 + \gamma \cdot \beta_2) + e_j^s + \upsilon_j \cdot \beta_2 (3)$$

The equation (3) implies that consumers are likely to exhibit different sensitivities in the searching stage from in the purchasing stage.

3 – 2 Model–free Evidence

In this section, I empirically test whether the consumer sensitivities to key attributes of Chromebooks are indeed different by using search rank and sales rank. Unlike the sales rank of products, there is no measure of search rank provided by Bestbuy.com. Although it is explained how view rank list is constructed above, it only shows the relative rank of products ‘within’ a focal product. Thus, as shown in Table 10, search scores of each product are calculated by using the concept ‘search popularity’ from Table1. Then integer search ranks are created based on products’ search scores.

As shown in Figure 7, all three graphs show similar patterns, implying that the relationship between the two ranks is robust to the definitions of search popularity. Table 11 is the regression results of different search ranks on products’ attributes. It implies the robustness of different definitions of search ranks. The blue dots in Figure 7 represent the laptop products. The black diagonal line is 45 angle line and products on those lines have the same sales and search rank. And the red line is the regression line of sales rank projected on search rank. The blue dots located above the black line are the products that are more popular in the purchasing stage than in the searching stage. Blue dots located below the black line are the products that are more popular in the searching stage than in the purchasing stage. I conjecture that the difference between sales rank and search rank of the same product derived partly from consumer’ s different sensitivities to key attributes in two stages and products’ different search cost.

Table 11. Constructing a search score, search rank, and sales rank

1. Search score

$$SearchScore_j = \sum_{l \neq j}^{J} Weight_l \times SearchPopularity_{jl}$$

where $SearchPopularity_{jl}$ is product j’ s popularity on l’ s view rank list and $Weight_l$ is the weight for the focal product k:

$$Weight_l = \frac{\#\ of\ total\ products + 1 - \ SalesRank_l}{\#\ of\ total\ products}$$

2. Search rank

   Search ranks of items are calculated by ranking their search score. Then, Inverse search ranks are calculated as follows:

$$Inverse\ Search\ Rank_j = \max_{j'}(search\_rank_{j'}) + 1 - \ search\_rank_{j'_j}$$

3. Sales rank

   Sales ranks of all items are averaged over a time period. Integer sales ranks are calculated using the average sales rank. Then, the integer sales ranks are inverted as follows:

$$\begin{aligned} &\mathit{Inverse\,Sales\,Rank}_j \\ &\qquad = \max_{j'}(\mathit{integer\_sales\_rank}_{j'}) + 1 - \ \mathit{integer}_{\mathit{salesrank}_{j'}}{}_j . \end{aligned}$$

Figure 7. Scatter plots of sales rank and search ranks

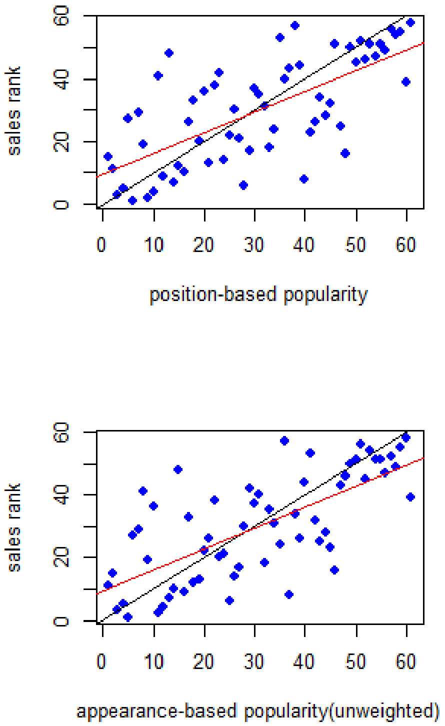


Table 12. Regressions of Search ranks under different definitions

Search rank Variables	Search rank (Appearance)		Search rank (Position)	
	Estimate	Std. err	Estimate	Std. err
Intercept	70.97(*)	31.98	69.92(*)	33.27
Log(Price)	−9.70(.)	5.59	−11.67(.)	5.81
Total rating	4.64	2.67	5.19(.)	2.78
Review #	−6.2e−04	6.1e−03	1.1e−03	6.3e−03
Two−in−one	−0.91	4.15	0.88	4.32
Screen size	0.41	1.37	0.84	1.42

Storage capacity		0.071	0.05	0.06	0.05
Ram		−1.14	1.88	−0.73	1.96
eMMC		21.35(*)	8.41	22.50(*)	8.75
Black color		−13.94(.)	7.82	−10.77	8.13
Acer		−10.82	7.51	−12.24	7.81
ASUS		−2.31	7.54	−0.68	7.85
Dell		7.94	7.63	6.43	7.94
Google		14.45(*)	12.9	15.95	13.52
HP		−15.44	6.91	−16.55(*)	7.19
Lenovo		−1.3e−03	8.17	−3.67	8.50
Position in listing page		−0.69(***)	0.13	−0.65(***)	0.14
R^2	obs	0.708	60	0.685	60

In order to test our conjecture, I estimate the following set of regression equations across products,

$$Inverse\,Search\,rank_j = \beta_0^s + X_j^f \cdot \beta_1^s + X_j^{sc} \cdot \beta_2^s + \varepsilon_j^s$$

$$Inverse\,Sales\,Rank_j = \beta_0^c + X_j^f \cdot \beta_1^c + X_j^{nf} \cdot \beta_2^c + \varepsilon_j^c$$

where X_j^f and X_j^{nf} are row vectors of product attributes available at a product listing page and a product−specific page, respectively. X_j^{sc} is a search cost−shifting variable, the position of products in product listing pages. Lastly, ε_j^s and ε_j^c are idiosyncratic errors in search and sales equations, respectively.

As we have seen from Table 11, the regressions of different definitions of search rank are fairly robust. In order to choose the definitions of search popularity for the rest of analysis in this paper, the following facts are considered; the length of view rank data in products from Bestbuy.com are relatively short, compared with the counterparts from Amazon.com as one can see in Kim et al.(2010, 2016) and appearance−based view rank inequality matrix can have some ties because of its discrete nature. Therefore, the appearance–based view rank inequality matrix can have less information than the position–based one. Therefore, view rank inequality matrix and search ranks based on position–based search popularity are utilized for the remaining parts of the paper.

Table 13. Regressions of Search rank and Sales rank

	Search rank (Position)		Sales rank	
	Coef.	Std. err.	Coef.	Std. err.
Intercept	69.92(*)	33.27	12..82	34.45
Log(Price)	−11.67(.)	5.81	−9.61	6.12
Total rating	5.19(.)	2.78	2.35	3.04
Review #	1.1e−03	6.3e−03	0.01(**)	0.006
Two−in−one	0.88	4.32	9.05(*)	4.39
Screen size	0.84	1.42	2.60(.)	1.43
Storage capacity	0.06	0.05	−0.07	0.05
Ram	−0.73	1.96	3.01	1.98
eMMC	22.50(*)	8.75	17.87(.)	9.13
Black color	−10.77	8.13	−9.66	8.30
Acer	−12.24	7.81	−17.73(*)	7.75
ASUS	−0.68	7.85	−21.13(**)	7.53
Dell	6.43	7.94	−2.04	8.22
Google	15.95	13.52	16.78	14.81
HP	−16.55(*)	7.19	−13.54(.)	7.37
Lenovo	−3.67	8.50	−6.65	8.99
Position in listing page	−0.65(***)	0.14	−	−
Price review score	−	−	6.82	11.70
Speed review score	−	−	38.36(*)	15.38
School review score	−	−	10.83	14.58
R^2	0.6845		0.6859	

From Table 12, one can notice the different preference levels of some attributes between two regression results. These results imply that consumers think the total rating of products as important while searching. It might be due to the reason that the total rating is the most related indicator of review score of which consumers cannot know the exact value while searching. The reduced form analysis in this section supports our conjecture that consumers have different sensitivities during the searching and purchasing stage. In the next section, I propose our empirical model that can explain these different consumer sensitivities by accommodating the product information gap between searching and purchasing stages.

4 Model

4 – 1 Utility and Empirical Specification

Assume the utility of a laptop j for a consumer i as follows:

$$u^c_{ij} = V^c_{ij} + e_{ij} = V^f_{ij} + V^{nf}_{ij} + e_{ij} = X^f_j \cdot \beta_{1i} + X^{nf}_j \cdot \beta_{2i} + e_{ij}$$

: purchase utility (4)

$$u^s_{ij} = \widetilde{V^c_{ij}} + e_{ij} = V^f_{ij} + \widetilde{V^{nf}_{ij}} + e_{ij} = X^f_j \cdot \beta_{1i} + \widetilde{X^{nf}_j} \cdot \beta_{2i} + e_{ij}$$

: search utility (5)

$$\text{where } e_j \sim Normal(0, \sigma^2)$$

$$and\ \widetilde{X^{nf}_j} = X^f_j \cdot \gamma + \upsilon_j,\ \upsilon_j \sim Normal\left(0, \Xi_{nf}\right)\ (6)$$

X^f_j is a row vector of costless attributes including price, brand dummies and other costless attributes. X^{nf}_j is a row vector of costly attributes, or review scores (speed, price, and school–use). I assume consumer heterogeneity follows a normal distribution, $\beta_i = [\beta_{1i}, \beta_{2i}] \sim \text{Normal}(\bar{\beta}, \Sigma)$, where Σ is a diagonal matrix. In Σ, diagonal elements corresponding to product price, brands, and reviews are nonzero and the other elements are fixed to zero.[3]

Since consumers are not able to observe X_j^{nf} before clicking-through, they form a belief of its distribution based on X_j^f. The equation (6) implies that consumers are already informed with the distribution of X_j^{nf} based on X_j^f. In other words, they are aware of its mean value, $X_j^f \cdot \gamma$ and variance, Ξ_{nf}. Therefore, before clicking-through, consumers have search utility with $\widetilde{X_{ij}^{nf}}$. Once the consumer decides to click the product, it will reveal its actual values of costly attributes. Therefore, the search process reveals V_{ij}^{nf} and e_j. The idea that set the different utilities is similar to Ghose et al. (2018) who also assume that consumers form a conditional belief on unknown values of attributes based on observable ones.

4 – 2 Optimal Sequential Search: Reservation Utility

The search cost is defined as

$$c_j = \exp(X_j^s \delta),$$

where X_j^s is a row vector containing variables affecting search cost including base search cost and j' s position in the product listing page. The position of j in the product listing page is included in the search cost part because Table 12 shows that it has a significant impact on search rank and Ursu (2018) also shows that the positions have an influence on users' click and transaction but not

[3] This is for avoiding overfitting the model with an excessive number of parameters. Only consumers' sensitivities that explain the model well under the heterogeneous specification are chosen.

on transaction conditional on click. Following Weitzman (1979), the solution to the sequential search problem can be characterized by a reservation utility rule. Define u^* as the highest utility among the searched products so far. Conditional on u^*, a consumer i' s expected a marginal benefit from the search of a product j is

$$\mathrm{B}_{ij}(u^*) = \int_{u^*}^{\infty} (u_{ij}^s - u^*) f(u_{ij}^s) du_{ij}^s \quad (7)$$

where $\mathbf{f}(\cdot)$ is the probability density distribution of u_{ij}^s. unlike Kim et al. (2016), u_{ij}^s exists in equation (7) instead of u_j. It is similar to the setting of Ghose et al. (2018).

The i' s reservation utility of product j, z_{ij} is defined as the utility level that makes a consumer indifferent between stopping and continuing searching for j. z_{ij} can be defined with regard to c_j as

$$\mathrm{B}_{ij}(z_{ij}) = c_j$$

Let $\overline{u_{ij}^s}$ be the mean and τ_{ij}^2 be the variance of the search utility u_{ij}^s. Based on the model settings, the mean and the variance of the search utility can be written as follows:

$$\overline{u_{ij}^s} = \mathrm{E}[u_{ij}^s] = \mathrm{E}\left(X_j^f \cdot \beta_{1i} + \widetilde{X_j^{nf}} \cdot \beta_{2i} + e_{ij}\right)$$

$$= X_j^f \cdot \beta_{1i} + X_j^f \cdot \gamma \cdot \beta_{2i} + \mathrm{E}(\upsilon_j) \cdot \beta_{2i} + \mathrm{E}(e_{ij})$$

$$= X_j^f \cdot \beta_{1i} + X_j^f \cdot \gamma \cdot \beta_{2i}$$

$$and$$

$$\tau_i^2 = Var[u_{ij}^s] = \mathrm{Var}\left(X_j^f \cdot \beta_{1i} + \widetilde{X_j^{nf}} \cdot \beta_{2i} + e_{ij}\right)$$

$$= \beta_{2i}' \cdot \mathrm{Var}(\upsilon_j) \cdot \beta_{2i} + \mathrm{Var}(e_{ij})$$

$$= \beta_{2i}' \cdot \Xi_{nf} \cdot \beta_{2i} + \sigma^2 \quad (8)$$

Here one can see from equation (8) that consumers have a heteroskedastic variance of search utility. Also, it makes sense that consumers who place a high value on users’ reviews are sensitive to their uncertainty, which is reflected in the equation (8). Moreover, consumers have a higher variance of search utility than that of purchase utility(σ^2), which is in line with the consumer learning literature’ s setting.

In order to calculate the reservation utility of every product, one takes the steps suggested in Ghose et al. (2018) as follows:

$$c_j = \mathrm{B}_{ij}(z_{ij}) = \int_{z_j}^{\infty} (u_{ij}^s - z_{ij}) f(u_{ij}^s) du_{ij}^s$$

$$= \left(1 - \Phi\left(\frac{z_{ij} - \overline{u_{ij}^s}}{\tau_i^2}\right)\right)\left(\frac{\phi\left(\frac{z_{ij} - \overline{u_{ij}^s}}{\tau_{ij}^2}\right)}{1 - \Phi\left(\frac{z_{ij} - \overline{u_{ij}^s}}{\tau_i^2}\right)}\tau_i^2 - \left(z_{ij} - \overline{u_{ij}^s}\right)\right)$$

Let $\eta_{ij} = \frac{z_{ij} - \overline{u_{ij}^s}}{\tau_i^2}$. Then, the equation can be rewritten as follows:

$$\frac{c_j}{\tau_i^2} = \mathrm{g}(\eta_{ij}) = \left(1 - \Phi(\eta_{ij})\right)\left(\frac{\phi(\eta_{ij})}{1 - \Phi(\eta_{ij})} - \eta_{ij}\right),$$

If one can solve $\eta_{ij} = \mathrm{g}^{-1}\left(\frac{c_j}{\tau_i^2}\right)$, then one can obtain z_{ij} from $z_{ij} = \eta_{ij}\tau_i^2 + \overline{u_{ij}^s}$. it has two key differences from the corresponding part of Kim et al. (2010; 2016). First, the derivation of Kim et al. (2010; 2016) shows that reservation utility has a linear relationship with the expected utility which is uniform across searching and purchasing stage whereas ours has a linear relationship with the expected search utility. Second, the reservation utility of Kim et al. (2010; 2016) is affected by homoskedastic utility variance while ours is by heteroskedastic variance.

The result above implies that the rank of the reservation utility is a one−to−one mapping with the product index. Hereafter, product index j is sorted as the decreasing order of reservation utilities. The next section will show some probabilities essential for the model inference. Kim et al. (2016) provide their mathematical explanation in detail, so this paper simply shows the formula of the probabilities.

4 − 3 Search and Choice Probabilities

1) The probability to search k

$$\pi_k = \Pr\left(\max_{l=1,\ldots,k-1}\left(V_l^p + e_l\right) < z_k\right)$$
$$= \prod_{l=1}^{k-1} \Phi_l(z_k - V_l), k > 1$$

2) The probability to choose j

$$\Pr(\mathrm{j}) = \sum_{K=j}^{J} \Pr\,(j, S_K)$$

where $S_K = [1, .., K]$ is ordered set such that if $z_k \geq z_l, then\ 1 \leq k \leq \mathrm{l} \leq \mathrm{K}.$

3) The joint probability that the jth ranked product is chosen from S_K

$$\Pr(j, S_K) = \int_{z_{K+1}-V_j}^{z_K - V_j} \prod_{l\neq j}^{K} \Phi_l(V_j^p - V_l^p + e_j)\phi_j(e_j)de_j + I(j$$
$$= K)(1 - \Phi_j\left(z_j - V_j^p\right))\pi_j$$

4) The probability to choose j conditional on searching option l

$$\Pr(j|l) = \frac{\sum_{K=\max(j,l)}^{J} \Pr(j, S_K)}{\pi_l}$$

5 Estimation and Identification Strategy

5 − 1 Pre−estimation

This subsection explains how I estimate γ and Ξ_{nf}. I use the Seemingly Unrelated Regression (SUR).

The equation (6) can be rewritten into matrix notation as,

$$\widetilde{X^{nf}} = X^f \cdot \gamma + \upsilon, \qquad \upsilon_j \sim Normal(0, \Xi_{nf}).$$

Note that $\widetilde{X^{nf}}$ is a matrix with rows consisting of product j(=1,⋯, J) and columns consisting of the costly attributes k(=1,⋯, K). Then, X^f is a matrix with the dimension of $\mathrm{J} \times \mathrm{K}'$ where K' is the number of the costless attributes, γ is with the dimension of $\mathrm{K}' \times \mathrm{K}$, υ is with the dimension of $\mathrm{J} \times \mathrm{K}$ and Ξ_{nf} is with the dimension of $\mathrm{K} \times \mathrm{K}$. Here I assume that υ_j is independent at the product level. That means consumer belief on the costly attribute of a product are common across consumers and does not affect his or her belief for other products. However, consumer belief on one costly attribute can be correlated with their belief on other costly attributes.

Based on this multivariate setting, the parameters γ and Ξ_{nf} are estimated in a SUR, in which the estimates are calculated using feasible generalized least squares (FGLS) in two steps. In the first step, separate ordinary least square regressions for each costly attribute are run. The residuals from each regression are used to estimate Ξ_{nf}. In the second step, γ is estimated by running GLS regression using the estimate of Ξ_{nf} . Then the estimated covariance Ξ_{nf} is used to calculate reservation utility as shown in section 4−2.

5 − 2 Main Model Estimation

The estimation of the joint model of search and choice closely follows Kim et al. (2016)' s procedures who also use aggregate−level data from Amazon.com. Since the aggregate indices in this paper are almost the same as those in Kim et al. (2016), the overall estimation procedure is similar. Thus, for estimation details, please refer to Kim et al. (2016). However, there is a difference in the estimation procedure that is the computation of reservation utility as shown in section 4−2. I use the mean and variance of search utility in order to calculate the reservation utility.

Kim et al. (2016) using random effects of some utility coefficients draw random values deviated from the mean value of coefficients in the utility. Each random value represents individual consumers' preferences. Using parameter estimates, they derive $\widehat{\mathbf{Pr}_i(j)}$ and $\widehat{\mathbf{Pr}_i(j|l)}$, aggregate them over consumers i, and use them as the estimates of market share and choice share of j conditional on l. Also, they compute $\widehat{\pi_{ik}}$, construct the predictions of commonality index and use them to predict pairwise view ranks. The parameter estimates are optimized by matching them against the aggregate measures from data and their standard errors are computed by using bootstrapping resampling technique(Efron and Tibshirani, 1986).

5 – 3 Identification

In this section, I discuss how parameters can be identified in this model. The parameters to be estimated in the main model include the mean utility, consumer heterogeneity parameters in utility function and the mean and product product–specific search cost parameters and variances of aggregate indices (equivalents of τ_V, τ_S, τ_C from section 3.2 in Kim et al. (2016)). The uncertainty of utility(σ^2) is fixed to be 1 for the identification, which is common in Probit–based choice models.

Before talking about how parameters are identified, I would like to explain how the change in parameters affects the search and choice probabilities. Mean utility affects the average search and choice popularity of products. As the value of mean utility coefficients increase, its probability to be searched rises. It also affects its probability to be purchased. An increase in search cost parameters reduces its reservation utility but do not affect the value of utility. Thus, it lowers the probability to be included in the consideration set. However, it does not affect the probability to be purchased once it is included in a consideration set.

Table 14. Relationship between parameters and probabilities

- Consumer' s preference for product j' s attributes ↑➔ V_{ij}^{c} and $\widetilde{V_{ij}^{c}}$ ↑ ➔ π_j, $\Pr(j,S_K)$, $\Pr(\mathrm{j})$ *and* $\Pr(j|l)$ ↑
- Search cost parameters ↑ ➔ z_j↓ , but V_{ij}^{c} and $\widetilde{V_{ij}^{c}}$ unchanged➔ π_j ↓, $\Pr(j,S_K)$ *and* $\Pr(j)$ ↓ *(marginally)*

From the relationships between variables and probabilities to be searched and purchased, it can be implied that if all products are searched and consumers have homogeneous tastes, the search rank of a product should be equal to its sales rank. This is because reservation utility will be totally up to the product utility. Black lines in Figure 7 are 45 angle and all products are on this line for this case. Therefore, the average search and choice popularity of products identifies the mean utility parameters. That means the mean utility parameters are identified by how correlated the variation in product popularity and the variation in product characteristics are and a strong positive correlation between search and sales ranks (being located near 45 angle degree) lead to more efficient parameter estimation in the joint model of search and choice.

However, as Figure 7 shows that a number of products deviate from the 45 angle line, products are likely to have a disparity between its sales popularity and search popularity. This gap can be explained by the search cost and heterogeneity. Since search costs do not affect the probability to be purchased once the product is put in the consideration set, it does not have as a strong impact on sales rank as the utility does. That means if there two products A and B, which have the same product attributes but different search cost, say, A has a higher search cost than B, then A should be located near B in y-axis and A be on the left side of B. Therefore, product-specific search costs are identified by the discrepancies between search and choice popularity.

The identification of consumer heterogeneity comes from the similarity between characteristics of a focal product and the products in its view rank list. For example, if there are consumers who prefer to search for products from the same brand than from the different brands, then it can be inferred that preferences for brands must be different across consumers. It can explain the slight deviation from 45 angle line in Figure 7.

The difference between its sales popularity and search popularity is also explained partially by the difference of search and purchase utility, and search utility variance. The expected values of costly attributes in search utility are governed by costless attributes not the actual values of costly attributes. Furthermore, Consumers have different search utility variances based on their preference of costly attributes. Therefore, combining all these factors make the disparity between purchase utility and reservation utility and hence, it explains the difference between sales popularity and search popularity.

6 Results

I investigate how well the proposed model fits the search and sales data patterns. This model achieves the good hit ratios of pairwise rank inequalities, in which the relative positions of two options in the actual and predicted rank data are compared: 85.5% for sales rank data and 80.2% for view rank data. These figures suggest that this model matches the search and sales patterns well. I also compare the proposed model' s performance to those of Kim et al.

(2016). To that end, I estimate the model of Kim et al. (2016) and calculate log-likelihood and hit ratios using the same dataset. The log-likelihood of Kim et al. is −20175, view rank hit ratio is 80.2%, and the sales rank hit ratio is 78.0%. Therefore, I conclude that the proposed model, which can be seen as the modified version of Kim et al. (2016) for the limited information product environment, shows better performance for the given empirical application.

Now I discuss the parameter estimates. The estimated brand intercepts have face validity: Google, the developer of Chrome and thus the most well-known brand for Chromebook, exhibits the highest mean brand coefficient of 0.28, Other brands show a similar level of negative coefficients as reduced-form estimation results show similar results in Table 12. The estimates show significant heterogeneity in brand preferences with an estimate of 0.62. For review-related estimates, Consumers prefer products with higher valence and volume of reviews. They also prefer products with positive reviews of actual users with regard to their subjective satisfaction with speed and school but they don’t seem to care much about satisfaction with a price. When it comes to design and feature, Consumers prefer Chromebooks of 2-in-1 design(0.08) and of colors other than black(−0.06). For the performance, consumers prefer Chromebooks of larger memory(RAM) (0.27). The estimates of storage are interesting. Consumers don’t prefer Chromebooks with larger capacity(−0.06) and like Chromebooks with eMMC type storage(0.57) than SSD. Considering that SSD is the more advanced type of storage than eMMC, the result is interesting. SSD is a faster storage type but more expensive than eMMC. Considering the average storage capacity of Chromebook (50.1GB in this sample) is smaller than that of other types of laptops, Chromebook does not require the storage capacity as large and fast as other types of laptops. Therefore, consumers might prefer ones with eMMC−type and smaller capacity.

With respect to search cost, consumers tend to click the products which are placed on the upper position in the product listing page. Similar results with regard to position effect are shown by Ursu (2018). Across products and consumers, the model predicts the mean and median search costs of \$0.26 and \$0.16, respectively. Using these parameters, I estimate a mean search set size of 51.

Table 15. Main model estimation results

	Variable	Estimates (s.e.)		Heterogeneity (s.e.)	
Utility	Log(Price)	−0.007	4.3e−04	0.05	2.4e−04
	Acer	−0.16	3.1e−02	0.62	5.6e−02
	ASUS	−0.24	4.8e−02	0.62	5.6e−02
	Google	0.28	7.6e−02	0.62	5.6e−02
	HP	−0.15	3.1e−02	0.62	5.6e−02
	Total Rating	0.16	6.0e−02		
	Review #	0.16	3.8e−02		
	2−in−1	0.08	2.3e−02		
	Screen size	−0.01	2.3e−02		
	Storage capacity	−0.06	1.6e−02		
	Ram	0.27	5.0e−02		
	eMMC	0.57	3.8e−02		
	Color Black	−0.06	1.1e−02		

	Speed (Review)	0.43	0.12	1.22	1.7e−06
	Price (Review)	−0.01	6.4e−02	1.22	1.7e−06
	School (Review)	0.26	0.16	1.22	1.7e−06
Search cost	Base search cost	−6.89	1.7		
	Position	4.80	4.6		
Aggregation	View rank	0.23	1.4e−03		
Error	Sales rank	0.004	3.7e−06		
	Conditional share	0.07	5.0e−05		
Log−likelihood: −20109, hit ratio: view rank− 80.2%, sales rank− 85.5% Number of pseudo−consumers: 1000					

7 Counterfactual Experiment

Using the estimated model parameters, I study the effects of different product information provision setting on consumers and producers. The current setting of product information revelation is limited information provided on product listing pages. Consumers have to click on the products in order to understand the values of costly product attributes (review scores). I simulate the counterfactual situation, where the values of review scores are provided on product listing pages. Therefore, Consumers only obtain horizontal matching values by clicking on the products. In the simulation, I use identical sets of 20000 consumers and simulate their search and choice decisions under limited information and full information scenarios.

Under the hypothetical full information scenario, consumers reduce their search set size by 3.9% on average. This is because consumers are motivated to search for products partly due to the uncertainty of costly attribute values. Since there is no uncertainty of costly attributes under the hypothetical situation, they decrease the search set size. Although their search set size differs according to the different scenarios, few consumers changed their choice decision. Almost all consumers choose the same products even under the counterfactual situation. Therefore, the revelation of costly attribute information gives them an increased surplus in that they buy the same products with reduced search efforts. Formally, I calculate the net surplus of a consumer with respect to a search set S_i which is defined as the highest utility in the search set less the total search cost incurred in the formation of i' s search set S_i:

$$NS(S_i) = \max_{j \in S_i}\{u_{ij}\} - \sum_{j \in S_i} c_{ij}.$$

Then I compute the aggregate change in the net surplus across consumers between the full information and limited information product environments. The total net plus increases by 3.19%.

I also examine the impact of product information provision on manufacturers' sales and revenue change. Consumers change their search and choice behavior in response to the change of information provision style. Compared with the limited information environment, most of the producers' market share and revenue increase. ASUS shows the largest increase in sales and revenue change and Google shows the negative change in sales and revenue. Sales change and revenue change of each brand under full information environment relative to limited information environment (Column 4 and 5 in Table 15) are highly negatively correlated with their total rating. Sales change and revenue change show correlation coefficients of −87.6% and −92.0% with total rating, respectively. This result implies that consumers might be less likely to search for products with a lower total rating because the total rating is one of the important indicators of review scores. However, total rating sometimes cannot fully reflect the review scores as ASUS does which is relatively low-ranked in total rating and highly ranked in average review score. In this case, consumers and manufacturers can take advantage of full information provision. Therefore, I want to suggest to manufacturers that they should post quantified review scores with respect to each attribute on product listing pages in order to boost their sales and revenues especially when their total rating is relatively low.

Table 16. Market change under full information environment

Brand	Total rating	Review score	# of products	Market share	Sales change	Revenue change
Acer	4.47	0.11	17	16.3%	17.1%	18.1%
ASUS	3.86	0.19	7	9.7%	31.3%	37.7%
Dell	4.51	0.09	8	7.3%	12.1%	11.8%
Google	4.68	0.14	4	20.7%	−1.1%	−1.5%
HP	4.6	0.19	12	12.1%	16.1%	14.1%
Lenovo	4.58	0.22	4	6.8%	11.7%	10.1%
Samsung	4.44	0.21	8	27.1%	13.8%	9.2%
Total			60		14.9%	10.4%

8 Conclusion

In this paper, I develop a joint model of search and choice that incorporates the different information sets consumers are presented with during search and choice. This model is applicable to when consumers can have a large search set and make a choice decision under a limited information environment. This model characterizes the uncertainty during the searching stage due to the unknown values of costly attributes of products within the framework of the costly search and choice decision driven by the same demand primitives. Uniform demand primitives during searching and purchasing stages allow for the parsimony of the model.

This paper makes the following contributions to the literature on consumer search and choice models. First, the model reflects the consideration that consumers are motivated to search products' information partly due to the uncertainty of products and they can solve the uncertainty, which is in line with the consumer learning literature. Second, the model reflects heteroskedastic uncertainty across consumers in a parsimonious way. Third, this paper takes advantage of a deep learning method in order to extract the structured features from reviews. Thus, this paper can give researchers motivations to use the methods of Artificial Intelligence.

I apply the proposed model to aggregate-level view and sales data from Bestbuy.com. Using the estimated demand parameters, I simulate the counterfactual situation where consumers can obtain costly quantified review scores on product listing pages. In this situation, both consumers and producers obtain increased surplus.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Chen, Y., & Yao, S. (2016). Sequential search with refinement: Model and application with click-stream data. Management Science, 63(12), 4345-4365.

Ching, A. T., Erdem, T., & Keane, M. P. (2013). Learning models: An assessment of progress, challenges, and new developments. Marketing Science, 32(6), 913-938.

Choi, H., & Mela, C. F. (2016). Online marketplace advertising. Available at SSRN.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 54-75.

Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. Marketing Science, 15(1), 1-20.

Gardete, P., & Hunter Antill, M. (2019). Guiding Consumers through Lemons and Peaches: A Dynamic Model of Search over Multiple Characteristics.

Ghose, A., Ipeirotis, P. G., & Li, B. (2018). Modeling consumer footprints on search engines: An interplay with social media. Management Science, 65(3), 1363-1385.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Gowrisankaran, G., & Rysman, M. (2012). Dynamics of consumer demand for new durable goods. Journal of Political Economy, 120(6), 1173-1219.

Honka, E. (2014). Quantifying search and switching costs in the US auto insurance industry. The RAND Journal of Economics, 45(4), 847-884.

Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2010). Online demand under limited consumer search. Marketing science, 29(6), 1001-1023.

Kim, J. B., Albuquerque, P., & Bronnenberg, B. J. (2016). The probit choice model under sequential search with an application to online retailing. Management Science, 63(11), 3911-3929.

Liu, L., Dzyabura, D., & Mizik, N. (2018, June). Visual listening in: Extracting brand image portrayed on social media. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

Liu, X., Lee, D., & Srinivasan, K. (2017). Large scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. Available at SSRN 2848528.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. Journal of Marketing Research, 48(3), 444-456.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. Marketing Science, 23(4), 579-595.

Song, I., & Chintagunta, P. K. (2003). A micromodel of new product adoption with heterogeneous and forward-looking consumers: Application to the digital camera category. Quantitative Marketing and Economics, 1(4), 371-407.

Stigler, G. J. (1961). The economics of information. Journal of political economy, 69(3), 213-225.

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. Marketing Science, 37(4), 530-552.

Weitzman, M. L. (1979). Optimal search for the best alternative. Econometrica: Journal of the Econometric Society, 641-654.

Appendix. Pre-estimation results

Table 17. R squared values

	R squared
Speed	0.374
Price	0.545
School	0.427

Table 18. The covariance matrix of consumer belief on costly attributes (Ξ_{nf})

	Speed	Price	School
Speed	0.017	0.0076	0.004
Price		0.022	0.005
School			0.015

초 록

본 연구에서는 제한된 상품 정보가 제공되는 환경에서 소비자가 상품을 검색하고 구매하는 행위를 설명하는 실증적 모형을 개발하였다. 본 모형에서 소비자는 상품 리스트 페이지에서 제품을 검색하는 과정에서 상품을 클릭 한 후에만 볼 수 있는 vertical attribute과 horizontal attribute에 대한 기대치를 가지고 있다(본 연구에서는 이를 costly attributes이라고 명함). Vertical costly attributes에는 상품의 몇가지 특성에 대한 리뷰 점수들이 포함되어 있다. 이 리뷰 점수들은 특정 상품 특징에 대한 실제 소비자 만족도를 계량화 한 것이다. 본 연구는 기존 연구에 다음과 같은 학문적 의의를 가진다. 첫째, 본 모형은 검색하기 전 단계에서 제품 효용에 대한 더 높은 불확실성을 가지고 있음을 반영하고 있다. 불확실성은 costly attribute에 대한 정보를 얻음으로써 어느정도 완화되는데, 이는 기존 소비자 학습 문헌과 일맥상통하는 바이다. 둘째, 본 연구는 모형을 복잡하게 만들지 않으면서, 상품을 검색하는 동안 소비자가 검색과정에서 이분산적인 효용을 가지는 것을 반영하였다. 셋째, 본 연구는 리뷰로부터 구조화된 변수를 추출하기 위해 딥러닝(deep learning) 모형을 활용하였다.

해당 모형은 베스트바이닷컴(Bestbuy.com)에 있는 크롬노트북 카테고리에 있는 검색 및 구매 데이터에 적용되었다. 그리고 현실적인 모수 추정치가 나왔고, Kim et al. (2016)와 비교했을 때, 해당 데이터에 대해 더 나은 적합도를 보여주었다. 그리고 추정된 모수를 통해 가상적인 상황에 대한 실험을 하였다. 해당 실험에서는 상품정보가 리스트 페이지에서 모두 제공되는 상황에서 소비자의 검색량과 제조업체의 시장 점유율 및 수익률이 어떻게 변하는지 보았다. 가상적인 상황에서 소비자는 −3.9%만큼 검색량을 줄이는데, 최종적으로 선택하는 상품은 거의 변화가 없었다. 그래서 변화된 정보제공 환경에서 소비자잉여 3.19%만큼 증가하였다. 그리고 대부분 제조업체의 시장 점유율과 수익률이 증가하였다. 더불어서 total rating이 비교적 낮고

리뷰 점수가 높은 브랜드가 상대적으로 더 큰 폭의 증가세를 보였다. 이 결과는 어떤 업체가 total rating이 상대적으로 낮다면, 상품 리스트 페이지에 상품 특성들에 대한 리뷰 점수를 게시하여 판매율과 수익율을 증가시킬 수 있음을 시사한다.

주요어 : 소비자 정보 검색, 소비자 학습, 제한된 상품 정보 제공 환경, 복잡형 인공신경망, 리뷰 데이터, 내구재
학　번 : 2018-29680